Fall 12-17-2021

mRNA-Sequencing Pipeline for Differential Gene Expression Analysis

Crystal Han
San Jose State University

mRNA-Sequencing Pipeline for Differential Gene Expression Analysis

A Project

Presented to

The Faculty of the Department of Computer Science

San José State University

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science in Bioinformatics

By

Crystal Han

December 2021

ABSTRACT

Endothelial cells (ECs) line the insides of blood vessels and play a key role in the coagulation and vascular repair system. Under normal circumstances, ECs are constantly expressing anticoagulants to prevent clots and maintain healthy blood flow. But when there is an injury to the vessel, endothelial cells become centrally involved in orchestrating the complex series of events that would lead to the clotting of the wound. Thus, ECs are dynamic and respond accordingly based on changes in their local environment. Unfortunately, endothelial cells can sometimes misinterpret cues from the environment and initiate coagulation when there is no vessel damage to repair, causing unwanted clots. This phenomenon was recently observed among astronauts who have spent prolonged time in a weightless environment. More research is needed to better understand how endothelial cells respond to microgravity conditions. This paper focuses on building an mRNA-sequencing pipeline for analyzing the gene expression of endothelial cells.

Keywords: Microgravity, Endothelial cells, mRNA-sequencing (RNA-seq), Transcription profiling

ACKNOWLEDGEMENTS

First and foremost, I want to thank Dr. Wendy Lee for being my advisor, for all of her help and guidance, and for constantly inspiring me through this process. I've learned so much from her and am lucky to be able to do my master's project with her.

I want to thank Dr. Anand Ramasubramanian for being on my committee and for the opportunity to work on his microgravity project, which I am excited to be a part of. I also want to recognize the rest of his team, including Dr. John Lee and their students.

I want to thank Dr. Philip Heller for being on my committee as well, and for all of his thoughtful feedback and insight.

I want to thank Alexis Torres for being the best teammate on the broader project, and for his help and support along the way.

I want to thank the rest of the Bioinformatics community at SJSU, including my fellow MSBI peers, for all of their support and encouragement.

Lastly, I want to thank my parents and sister for always being there for me. I would be lost without them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

Space exploration is an exciting and growing field, and there are ambitious dreams to send humans to the moon, Mars, and beyond. However, it is also not unknown that spaceflight takes a heavy toll on the human body and there can be significant, lasting adverse effects. In a recent 2019 study that followed a cohort of crew members to the international space station (ISS), one particular health risk was brought to the forefront [1]. Of the 11 astronauts in the study, it was found that six of them demonstrated stagnant or reversed flow in the internal jugular vein (IJV), which is a risk factor for blood clots, and one member actually developed an occlusive IJV thrombus during spaceflight. It is clear that we need to have a better understanding of how weightlessness and fluid shift in space impact cardiovascular health and the risk of blood clot formation.

The goal of this project is to construct an mRNA-sequencing (RNA-seq) pipeline for analyzing the gene expression of endothelial cells (ECs) and, in particular, how venous EC expression changes under microgravity and flow conditions. This is part of a larger study funded by NASA Space Biology entitled "Thrombosis in Microgravity," whose goal is to investigate the effects of microgravity and fluid flow rate on IJV thrombosis by conducting experiments using human umbilical vein endothelial cells (HUVECs). Specifically, the study will focus on the analysis of the EC transcriptome and differential expression of genes under the following four conditions presented in Table 1.

No flow, 1g (control)	Normal flow, 1g (protective)
No flow, microgravity	Normal flow, microgravity

Table 1. Experimental conditions for endothelial cell culture

Unfortunately, due to time constraints, experimental results from the study were not available yet at the time of this writing, so public data from published papers will be used to construct and validate the pipeline, with the understanding that it can be easily transferable to the experimental data once it is ready.

1.1 Endothelial cells and the coagulation system

Endothelial cells line the insides of blood vessels and play a central role in the vascular repair system, including the formation of clots to stop bleeding upon injury. Blood clots, or thrombuses, are formed from a combination of platelets and a mesh of fibrin proteins. During clotting, the Von Willebrand factor helps adhere platelets to the wound site. At the same time, an enzyme called Factor X is activated to form Factor Xa, which catalyzes the formation of thrombin. Thrombin plays a role in platelet aggregation and fibrin formation, effectively forming a blood clot [2].

Under normal circumstances, ECs are constantly expressing antiplatelets and anticoagulants to prevent clots and maintain healthy blood flow. One of the key genes involved in this process is *TFPI* (Tissue Factor Pathway Inhibitor), which codes for a protease inhibitor that inhibits Factor Xa. Other anticoagulants expressed by the endothelial cells include thrombomodulin, *EPCR*, and heparin-like proteoglycans [2].

There are many external factors that can trigger the endothelial cells to switch from an anticoagulation state to thrombus formation. This includes physical damage to the vessel wall, but also more subtle shifts in the vascular environment, such as changes in oxygen availability, pressure, and shear stress [2]. The Virchow triad describes three factors that can increase the risk of venous thrombosis, one of which is slowed blood flow or stasis. The lack of flow causes the

pooling of procoagulant factors, such as thrombin, which promotes thrombosis [3]. Stasis is a cause for concern in microgravity environments like during spaceflight.

1.2 RNA-seq analysis tools and workflow

One way to investigate how endothelial cells respond differently to various external conditions is to perform RNA-seq analysis. The presence of mRNA indicates which genes are being expressed at a given moment. By quantifying how much mRNA is present and comparing their abundance between control and treatment conditions (e.g., normal vs. microgravity), we can see how certain genes become differentially expressed as a result of a changed condition.

Once the mRNA have been sequenced in the lab, the RNA-seq analysis process usually consists of quality assurance steps, mapping the sequenced reads to a reference genome, and performing differential expression analysis. There are multiple tools available that can be used to conduct each part of this process. Put together, they form the analysis pipeline.

1.2.1 Assessing read quality

The raw sequenced mRNA reads are usually provided in a FASTQ format, which stores both the nucleotide sequence and the associated quality score for each base. The first step is to evaluate the quality of the sequencing reads and identify any potential issues that may affect the downstream analysis. FastQC is a quality control tool that takes FASTQ files as input and generates summary metrics about the reads, including the per base sequence quality, sequence duplication levels, and adapter content [4]. The report is outputted as an HTML file, which displays the results as statistics and summary tables and graphs. FastQC was chosen for its ease of use, simplicity, and easy to interpret output.

Based on the quality assessment, measures can then be taken to address any problematic issues, such as trimming adapters and low-quality sequences as well as removing duplicate reads.

1.2.2 Trimming adapters and low-quality sequences

During sequencing, the accuracy of base calls often drops towards the end of the read due to the accumulation of errors that may occur along the way [5]. This would be indicated by the quality score associated with each base in the FASTQ file. Another issue that can occur during the sequencing process is that for shorter sequences, the length of the fragment may be less than the number of bases being sequenced, thus the sequencing can continue into the adapters that are ligated to the ends of the fragment [6]. As a result, these reads can end up containing the adapter sequences.

Trimmomatic is a flexible trimming tool that is able to address both low-quality reads and adapter contamination [7]. One of its features include using a sliding window to detect where the average quality of the bases fall below a certain threshold, and cutting the read at that point. It can also trim user-provided adapter sequences from the reads by aligning the two and clipping the reads if the mismatch is below a specified threshold. In addition, Trimmomatic has several other trimming options, including dropping reads below a certain length, which may be useful for more robust alignments in the subsequent mapping step.

1.2.3 Mapping reads to the genome

Once the quality of the sequencing reads have been assured, the next step is to map them to a reference genome to determine where the sequences originated from. Since the reads come from RNA sequences and not DNA, it is crucial to take into account any splicing that may have occurred. Mature mRNA will have undergone several processing steps post-transcription, which includes the removal of introns, so the resulting reads will not be able to be mapped directly to the genome without consideration of the splice sites.

Luckily, there are various tools available that are designed to be able to map RNA reads back to non-contiguous parts of the genome. STAR is the splice-aware aligner chosen for this project because of its speed and accuracy [8]. STAR works by attempting to map each read as far as it can to a contiguous portion of the genome, then repeating the process but only for the remaining unmapped portion of the read, thus effectively cutting down on runtime, as compared to other algorithms that attempt to find all possible full read matches. Although STAR is fully able to detect splice junctions on its own, its accuracy can further be improved by providing a gene annotation file that identifies the specific splice sites. STAR also provides the capability to generate genome index files prior to mapping, which helps save time and memory during the mapping process.

The output is an alignment file that stores each read with its alignment to the reference genome, along with additional information like mapping quality scores. The file can be outputted as Sequence Alignment Map (SAM) format, or the compressed binary form, Binary Alignment Map (BAM) format, which can save disk space. A read can either be mapped uniquely to one location of the genome, mapped to multiple locations if it is ambiguous where it came from, or unmapped if no alignments can be confidently determined. STAR provides options to specify whether and to what extent to retain unmapped reads or multi-mapped alignments in the SAM or BAM file.

1.2.4 Removing duplicate reads and assessing alignment quality

During the RNA-seq library preparation process, the sequences typically undergo PCR amplification to produce enough copies to be detectable by the sequencing machine. But due to PCR biases, not all sequences may be amplified equally [9], which can lead to misleading results later down the road during read quantification, as overrepresented sequences can falsely give the

impression of higher expression levels. Another potential source of duplication can occur during the sequencing process itself, where a single clonal cluster being sequenced might be misinterpreted as multiple clusters, resulting in what are known as optical duplicates. It is important to account for these errors and biases in order to accurately quantify the reads.

Picard is a collection of command-line tools for working with sequencing data in formats such as SAM or BAM [10]. It includes a program called MarkDuplicates that is capable of flagging and removing duplicates by checking for reads that align to the same positions of the genome, which is unlikely to occur by chance since sequence fragmentation was random, and retaining the read with the highest base quality [11]. An advantage of removing duplicates at this stage of the pipeline using the aligned reads instead of directly comparing the raw reads for exact matches is that it is not limited by any sequencing errors that may have occurred.

In order to use MarkDuplicates, the input alignment file must be sorted. If it is sorted by coordinates, the tool will not mark the unmapped mates of mapped records or any secondary or supplementary alignments as duplicates, whereas it can if the file was sorted by query name. Picard includes a tool in its collection called SortSam for sorting SAM and BAM files [12]. It comes with the option of generating an accompanying index file that can speed up file lookup. Sorting and indexing the alignment file is useful to make accessing the contents of the file faster for downstream tools. In addition, Picard also provides several programs for collecting quality metrics about the alignment files. Its CollectMultipleMetrics tool is able to call on these programs in a single step to collect multiple types of metrics [13].

1.2.5 Quantifying reads

Once the quality of the alignments have been assured, the reads can be quantified to determine how many reads are mapped to each gene in the reference genome. featureCounts is a

quantification tool that takes SAM or BAM files as input, as well as the corresponding gene annotation file, and generates the raw counts of reads mapped per genomic feature [14]. For paired-end data, it is able to count each pair of reads once as a single fragment rather than separate reads. Internally, featureCounts will sort the alignment file by query name such that the paired reads always follow one another consecutively, so the input file does not necessarily have to be pre-sorted.

featureCounts works by quantifying the alignments between the reads and the genomic features provided in the annotation file. It can generate counts for features such as exons or broader, meta-features like genes, which would be the interest of this project. When quantifying to the meta-feature level, only the number of overlaps with each meta-feature is counted. For example, even if a read spans multiple exons, it would only contribute one count towards the overarching gene. Multi-mapping reads that were mapped to more than one location in the genome are not counted at all by default. In addition, if a read overlaps multiple genes, featureCounts by default does not count that read as there is not enough confidence to attribute it to any one gene. However, in the case of paired-end data, if there is ambiguity involving multiple genes, featureCounts further checks to see if there is one gene that overlaps both reads in the pair while the other overlaps just one read, in which case there is enough confidence to count the read towards the former.

Compared to other quantification tools, featureCounts works particularly well for paired-end data because of its ability to potentially resolve ambiguity for fragments overlapping multiple genomic features. featureCounts is also much faster and less memory intensive than similar tools [14].

1.2.6 Differential gene expression analysis

After the quantification step, gene-level differential expression (DE) analysis can be performed to identify any genes that are expressed in significantly different amounts between samples under different conditions. DESeq2 is an R package that takes raw counts data as input and provides functions for detecting differential expression [15]. Raw counts data from RNA-seq experiments will typically have a heavily right-skewed distribution, where a majority of the genes have very few counts, while few genes are highly expressed. In addition, the variance across samples for each gene tends to become larger than the mean, particularly for genes with high mean expression levels. These criteria make the Negative Binomial distribution an ideal approximation for RNA-seq counts, which is what DESeq2 uses to model the data [16].

For a gene to be considered differentially expressed between treatment and control groups, the mean expression level should be significantly different between the groups, while the variation within a group is minimal. Internally, DESeq2 normalizes the raw counts to account for the differences in sequencing depth and RNA library composition to make the samples comparable with each other. It is not necessary to account for gene length since comparison is not done across genes [16]. To estimate the variation in the data, DESeq2 calculates the dispersion for each gene, which is a better measure than the variance. As mentioned previously, the variance tends to be lower for genes with low mean expression, making it easier for genes with low counts to be identified as differentially expressed. On the other hand, the dispersion accounts for the variation at different mean expression levels [16].

Once the data is optimized and the model is fit, DESeq2 performs hypothesis testing using the Wald test to determine whether there is a significant difference in the mean expression between the treatment and control groups for each gene. The difference in expression is

measured as a log2 fold change (LFC), where the null hypothesis assumes a LFC of 0. Among other statistics, DESeq2 returns the LFC, associated p-value, as well as the adjusted p-value for each gene. The adjusted p-value, corrected by the Benjamini-Hochberg method, represents the false discovery rate (FDR) of identifying a gene as differentially expressed [16]. It should be used during multiple hypothesis testing when multiple genes are tested at the same time.

1.2.7 Gene set enrichment and pathway analysis

In addition to identifying individual DE genes, gene set enrichment analysis (GSEA) can be conducted to detect whether groups of genes belonging to known pathways are significantly enriched or depleted. Sometimes, individual genes might not be determined as differentially expressed according to some significance threshold, but put together, genes from a common pathway can show an overall significant difference in expression at the gene set level [17].

To perform GSEA, the genes must first be ranked according to their differential expression, where genes at the top and bottom of the list are those that are highly upregulated or downregulated, and genes in the middle are those without significant difference in expression. Common metrics for ranking include using the shrunken LFC, Wald test statistic, or signed p-value. For a gene set to be considered enriched or depleted, the genes belonging to that set should be found mostly near the top and bottom of the ranked list, rather than randomly distributed across the list. An enrichment score (ES) is calculated to measure how well the genes of a gene set are aggregated near the extremes of the list, with a positive value indicating they are mostly distributed near the top of the list and a negative value indicating they are mostly at the bottom. The normalized enrichment score (NES) accounts for gene set size and other factors to make it comparable across gene sets. An associated adjusted p-value is also calculated to represent the FDR of a gene set being identified as enriched [18].

fgsea is an R library for performing GSEA [19]. It takes a ranked list of genes and collection of gene sets as input and outputs the ES, NES, and adjusted p-value, among other statistics. It can also generate enrichment plots and summary tables for being able to quickly visualize results.

1.2.8 Aggregating log files

During each step in the pipeline, there may be output or log files produced by the various tools for each sample. This can quickly amount to a significant number of files, making it difficult to manually check on each one. MultiQC is a results aggregation tool that is able to compile the output files from multiple tools into a single HTML report [20]. It is able to recognize and parse files from a wide variety of common Bioinformatics tools and display summarized statistics and interactive plots. For example, FastQC, STAR, and featureCounts are among the many tools supported by MultiQC. The resulting aggregated report makes it easy to quickly assess the outputs across all samples, from read quality to mapping and quantification results.

1.2.9 Workflow management system

To help manage the workflow, Snakemake was used as the Bioinformatics pipeline framework [21]. Having a framework not only allows the user to run the complete set of tools together in a pipeline, but also provides additional options and flexibility, such as the ability to run steps in parallel and easily swap out files and settings using a configuration file. In addition, if the pipeline fails at a particular point or change is made to only one part of the pipeline, the system would be able to detect and rerun only the affected portions, thus saving time and increasing efficiency.

With Snakemake, each step in the pipeline is defined as a rule that specifies the expected inputs and outputs as well as the shell command or script for running that step. For each rule, an isolated software environment can also be defined to be able to use specific versions of the tools involved at that step. Newer versions of Snakemake use mamba by default to install conda packages, which is a faster and more robust reimplementation of the conda package manager. This helps make Snakemake workflows highly scalable and reproducible.

To run the pipeline, the Slurm Workload Manager was used to schedule jobs and allocate resources on the San Jose State University (SJSU) High Performance Computer (HPC) cluster [22]. Snakemake also has the capability to integrate the execution of workflows on a cluster.

II. DATA AND METHODS

2.1 Data

To test and validate the pipeline, RNA-seq data from two recent papers were used. The first dataset was specifically chosen to be one that is studying the effects of microgravity, as it relates to the scope of the original project, while the second is a more general dataset. Furthermore, the datasets were selected to contain both mRNA and miRNA sequencing data, since the original project will be studying both, to allow for exploratory analysis of the interaction between mRNA and miRNA. This limits the available datasets to only one mouse dataset for the first paper, which is why mouse data was used to test the pipeline.

2.1.1 Paper #1

The first dataset comes from a 2021 paper entitled "Cerium oxide nanoparticle administration to skeletal muscle cells under different gravity and radiation conditions" by Genchi et al. that was published in ACS Applied Materials & Interfaces [23]. The mRNA-sequencing reads data is available in NCBI’s Gene Expression Omnibus (GEO) database

(GSE165565), along with the raw counts data outputted from the quantification step, which can be used to validate the project pipeline results.

In this study, the authors aimed to understand the effects of cerium oxide nanoparticles ("nanoceria") on the transcriptome of skeletal muscles under various gravity and radiation conditions. Table 2 shows all the experimental conditions of the paper, but this project will only focus on the samples without nanoceria treatments (i.e., experimental class type A, C, E), as that is outside the scope of this project.

experimental_class_type	regime	treatment
A	in space without gravity	without nanoceria
B	in space without gravity	with nanoceria
C	in space with gravity	without nanoceria
D	in space with gravity	with nanoceria
E	on land	without nanoceria
F	on land	with nanoceria

Table 2. Experimental conditions for Paper #1

C2C12 mouse myoblasts (ATCC CRL-1772) were used in this study, with three biological replicates for each experimental class type. The cell cultures for class A and C were sent on board the International Space Station (ISS), whereas those for class E remained on Earth. Further, the samples for class C underwent centrifugation to simulate 1μ gravity on Earth, while still receiving the exposure of cosmic radiation in space. This yields three possible pairwise comparisons to explore the effects of microgravity and space radiation (A vs. E), only radiation (C vs. E), and only microgravity with a background of radiation (A vs. C). This project will focus on the two related to microgravity.

Total RNA extraction was performed using the MirVana PARIS kit (Ambion AM1556), and mRNA was isolated via poly-A tail selection and the sequencing libraries were prepared using the NEBNext Ultra RNA Library Prep Kit for Illumina. The samples were sequenced on the Illumina HiSeq 2500 System using a 2x150bp paired-end configuration.

2.1.2 Paper #2

The second dataset comes from a 2021 paper entitled "Transcriptional Network Analysis Reveals the Role of miR-223-5p During Diabetic Corneal Epithelial Regeneration" by Zhang et al. that was published in Frontiers in Molecular Biosciences [24]. The mRNA-sequencing reads data is available in NCBI's Gene Expression Omnibus (GEO) database (GSE180490).

In this study, the authors aimed to understand the mechanisms by which hyperglycemia (high blood glucose) leads to a loss of corneal epithelial regeneration function. Six-to-eight weeks old male C57BL/6 mice were used in the study, with half of them given streptozotocin (STZ) injections for five days to induce Type 1 diabetes. After 16 weeks from the final injection, the diabetic mice with a blood glucose level of over 25 mmol/L were selected. Three biological replicates were used each for the treatment and control groups.

To obtain the samples for RNA-sequencing, the corneal epithelium of the mice were first removed, then samples of the regenerative corneal tissues were collected after 24 hours of the injury. The mRNA library was prepared using the KAPA Stranded RNA-Seq Library Prep Kit and sequenced on the Illumina NovaSeq 6000 System using a 2x150bp paired-end configuration.

2.2 Methods

A summary of the workflow for the RNA-seq pipeline is shown in Figure 1, along with the tool versions below.

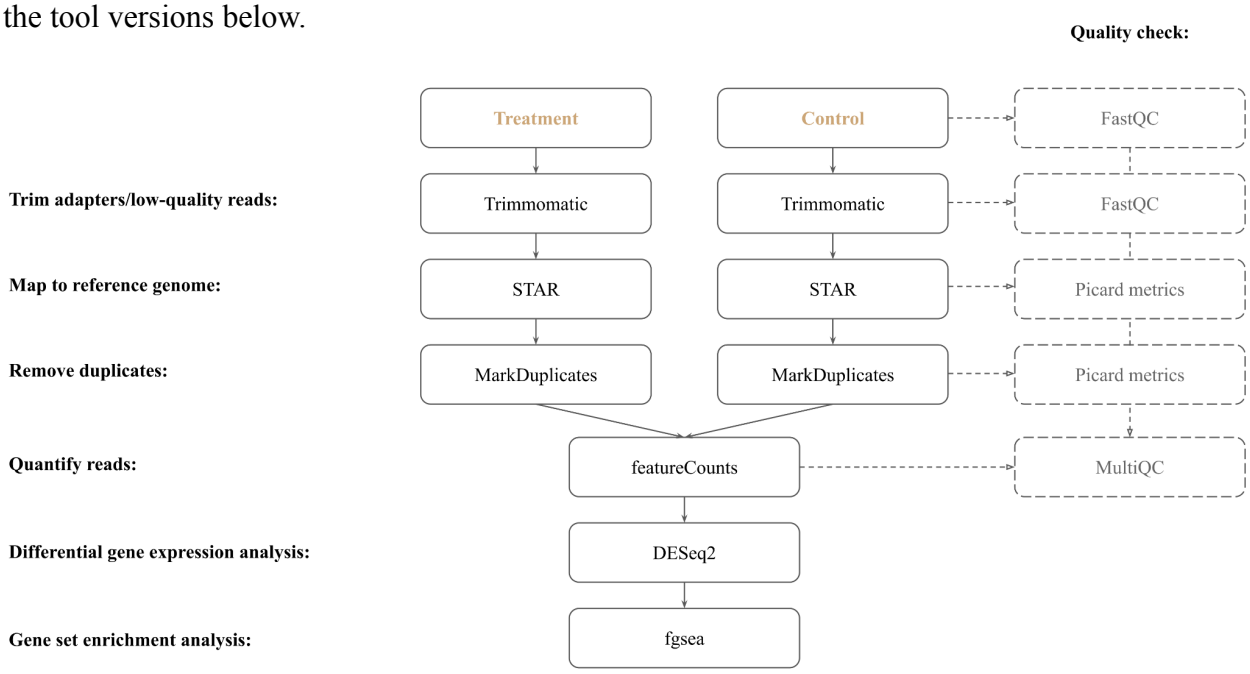

Figure 1. Flowchart for the RNA-seq pipeline workflow

- FastQC (v.*0.11.9*)
- Trimmomatic (v.*0.36*)
- STAR (v.*2.5.2b*)
- Picard tools (v.*2.23.0*)
- featureCounts (v.*1.5.2*)
- MultiQC (v.*1.10.1*)
- R (v.*4.1.0*)
- DESeq2 (v.*1.34.0*)
- fgsea (v.*1.20.0*)
- Python (v.*3.8.12*)
- Snakemake (v.*6.10.0*)

The full pipeline was run on the SJSU HPC using Snakemake. There are 72 compute nodes available on the HPC. Each node has 2 CPUs, and there are 14 cores per CPU for a total of 28 cores per node. There is also 128GB of RAM per node. In an attempt to reproduce the paper's results, the same tools and versions of the tools used in the first paper were used in this pipeline, though this was limited to what was presented in the paper. Since conda environments were used for each rule, this can be easily updated as needed.

The raw mRNA-sequencing reads for both papers were downloaded from the GEO database, with a total of nine samples from the first paper and six samples from the second. Because these are all paired-end data, there were two FASTQ files for each sample. Both the forward and reverse reads for each sample underwent quality assessment using FastQC. Among other metrics, the adapter content was checked to see what sequences need to be trimmed.

Trimmomatic was used to trim the adapters and low-quality sequences. FastQC identified the adapters to be the Nextera Transposase Sequence and Illumina Universal Adapter, and the respective sequences were obtained from the Trimmomatic GitHub repository for trimming [25]. In addition, a Phred quality score threshold of 20 was used to clip the reads where the quality starts to drop within a sliding window of 4 bases. A Phred score of 20 corresponds to a 1% error rate, which is considered acceptable for most purposes [26]. A minimum read length of 20bp was also used to filter out short reads that can cause ambiguity during the mapping process. The trimmed FASTQ files were again evaluated using FastQC to ensure the trimming generated the expected results.

Once the trimmed reads were satisfactory, they were mapped to the reference genome using STAR. Following the first paper, the mouse reference genome GRCm38.p6 was used. The primary assembly genome FASTA file and corresponding gene annotation GTF file were

obtained from Ensembl for the first paper [27] and UCSC Genome Browser for the second [28], as that is what produced the most consistent results when compared to the respective papers. The reference genome and annotation files are specified in the configuration file of the Snakemake pipeline, so it can be easily swapped out for each run.

The reference genome was first indexed using STAR. Since STAR is a splice-aware aligner, it simultaneously constructs a splice junctions database given the annotation file. To optimize the performance of the mapping, the STAR manual recommends specifying the maximum read length minus one as the number of bases around the junction sites to be used while creating the database. Since the data from both papers used a 2x150bp configuration, the reference genome was indexed using an overhang of 149. After mapping, the alignment files were outputted as BAM files to save disk space.

The BAM files were then sorted by coordinate position and indexed. Coordinate-sorting was used over sorting by query name because it is the preferred ordering for many visualization tools like the Integrative Genomics Viewer (IGV). As mentioned previously, MarkDuplicates will not mark the unmapped mates of mapped records or any secondary or supplementary alignments as duplicates when the file is sorted by coordinates. However, this does not have an effect on the subsequent quantification step because featureCounts does not by default look at unmapped reads or multi-mapped reads. Additionally, featureCounts will internally sort the alignment file by query name, so it is not required to provide the input sorted as such.

When quantifying reads, featureCounts also considers the strandedness of the reads. Depending on how the sequencing libraries were prepared, the forward and reverse reads may be marked to come from a specific strand of the DNA. The data from the first paper was determined to be unstranded, while the data from the second paper was reversely stranded. The

infer_experiment.py script from RSeQC was used to confirm the strandedness of the libraries [29].

Differential expression analysis was conducted using DESeq2 to identify genes that are significantly upregulated or downregulated between the treatment and control groups. An adjusted p-value of 0.05 was used as the significance threshold for analysis of both papers' data. In addition, the first paper used an absolute LFC of above 1 as a criteria for identifying DE genes, and the second paper used an absolute LFC of above 0.585 as well as a mean fragments per kilobase per million (FPKM) greater than 0.5 as their criteria, so these were applied to the respective analyses of the data as well. For the first paper, the raw counts file from their quantification step was available for download, so that was also obtained and run through the DE analysis to use as comparison.

For GSEA, the Molecular Signatures Database (MSigDB) provides nine major collections of gene sets, including the Hallmark gene sets for common biological pathways and the Gene Ontology (GO) gene sets, with the more specific Biological Process (BP), Cellular Component (CC), and Molecular Function (MF) subcomponents. These gene sets are available in R format for mouse data from Walter and Eliza Hall Institute (WEHI) [30]. Because these gene sets use Entrez ID's instead of Ensembl ID's, conversions between the gene ID's were obtained using the R package biomaRt [31]. Conversions between Ensembl ID and gene symbols were also obtained to be able to compare with the second paper's results, as those were presented using gene symbols. The Wald test statistic was used to rank the genes for the analysis. An adjusted p-value of 0.25 was used as the significance threshold for identifying enriched gene sets, as that is generally considered acceptable for exploratory GSEA [18].

III. RESULTS

3.1 Paper #1

Raw RNA-seq data from nine samples were downloaded for the first paper by Genchi et al. The total number of paired-end reads as well as the experimental condition for each sample is shown in Figure 2.

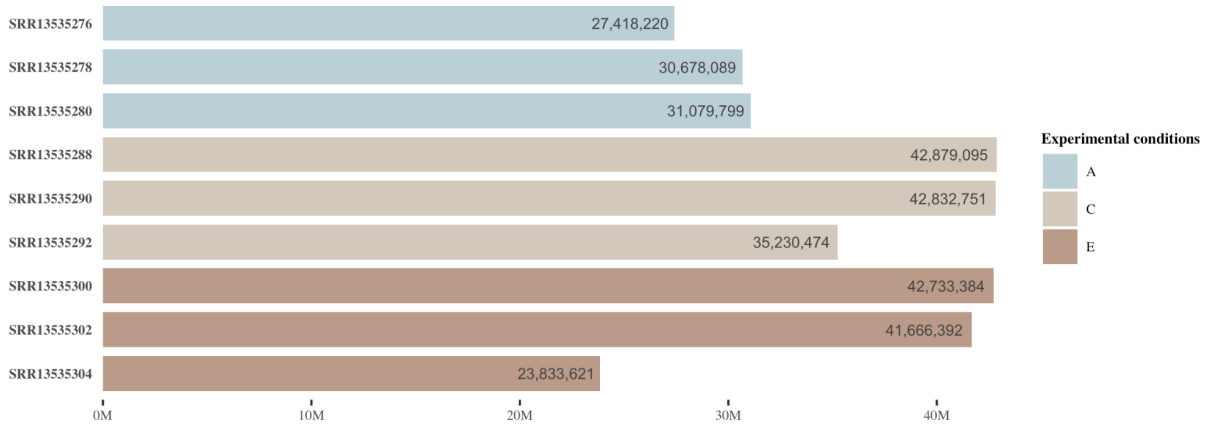

Figure 2. Raw read counts per sample, Paper #1

3.1.1 Trimming results

For the first paper's data, FastQC identified the adapter sequence as Nextera Transposase Sequence, and this was successfully removed using Trimmomatic, as seen in Figure 3. After removing low-quality sequences and short sequences, over 90% of the reads still remained in all samples, as seen in Figure 4.

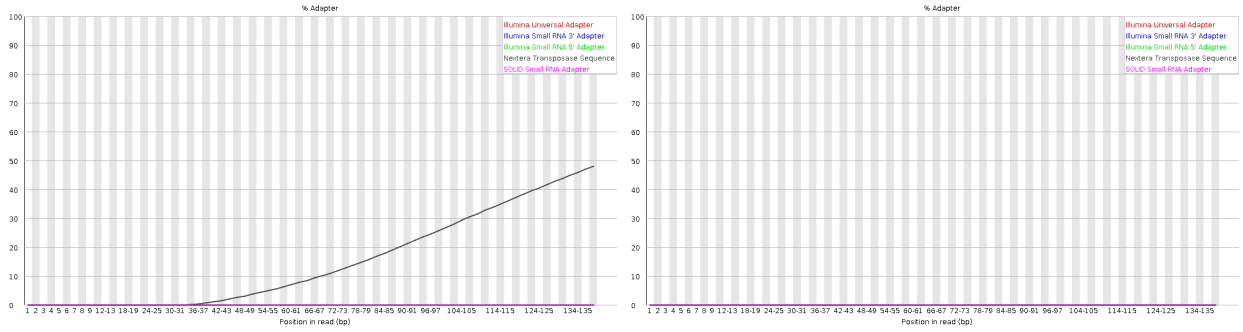

Figure 3. Adapter content before and after trimming, Paper #1

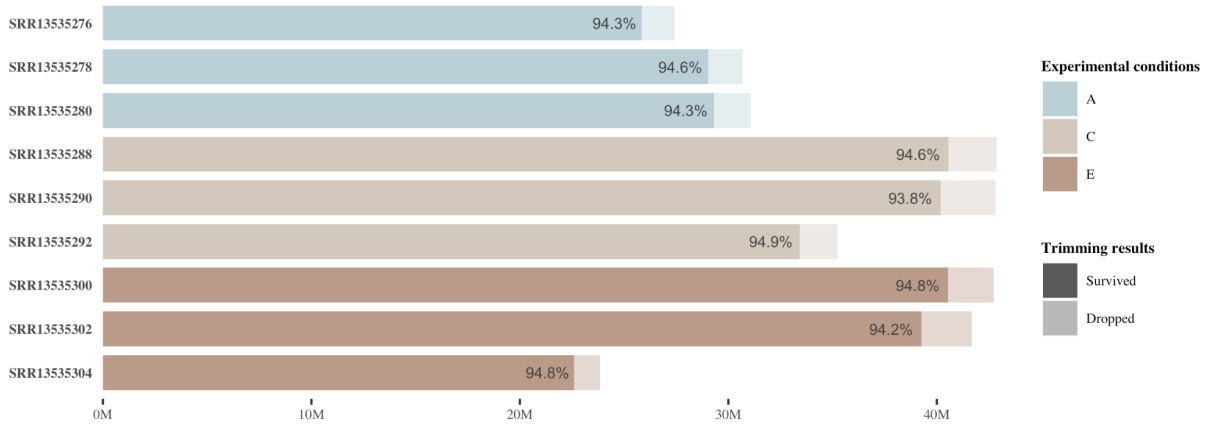

Figure 4. Trimming results, Paper #1

3.1.2 Mapping results

As seen in Figure 5, between 40-70% of the reads that survived the trimming step were uniquely mapped to the reference genome, which corresponds to at least 9M reads in each sample. There is also a relatively high proportion of reads that were unmapped, over 50% in some cases. The unmapped reads were largely classified by STAR as being too short, which by default is defined as alignments where less than two thirds of the read is mapped. It is possible to adjust this criteria to something less stringent to increase the number of mapped reads, however, this may lead to more inaccurate mappings, so that was not done. Furthermore, there are enough reads remaining to proceed confidently with the rest of the downstream analysis.

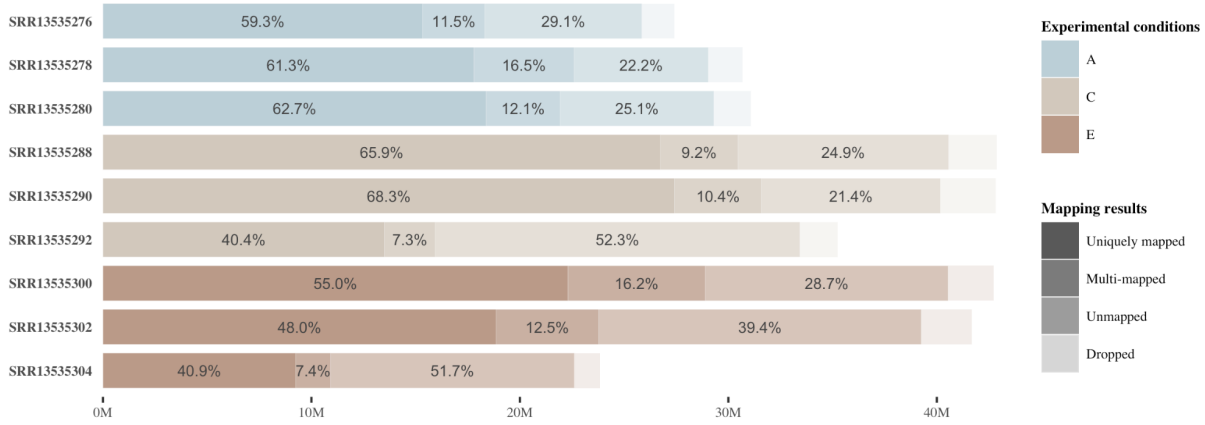

Figure 5. Mapping results, Paper #1

3.1.3 Deduplication results

Figure 6 shows the percentage of mapped reads that survived after removing duplicates. Only mapped reads are considered by MarkDuplicates because duplicate removal is done by comparing the start and end positions of the alignments. The number of mapped reads include reads that are uniquely mapped and reads that are multi-mapped by STAR.

For example, there are approximately 18M mapped reads for sample SRR13535276, over 73% of which are identified and removed by MarkDuplicates as duplicate reads.

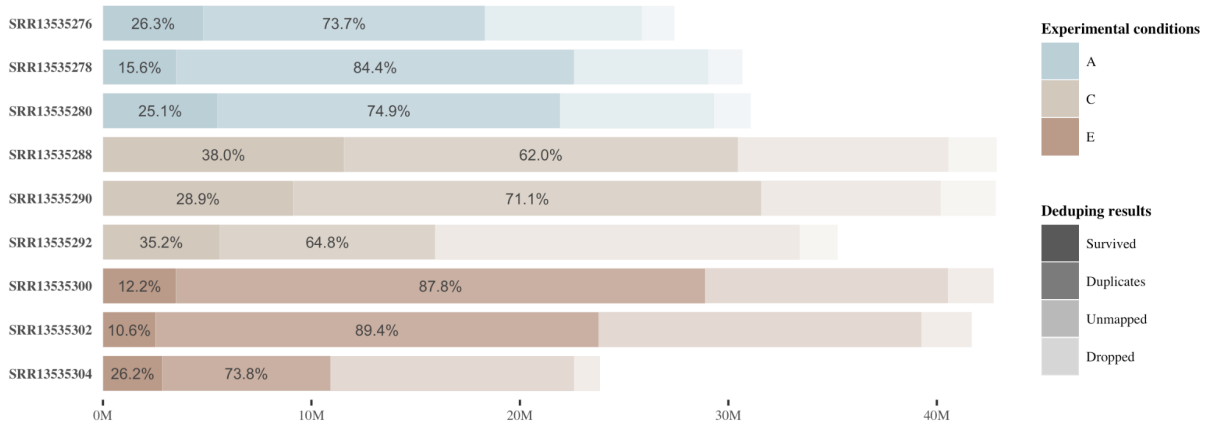

Figure 6. Deduplication results, Paper #1

3.1.4 Quantification results

Figure 7 shows the final number of read pairs that were assigned to a gene by featureCounts, relative to the original number of raw reads that each sample started with. This is the total number of reads that made it to the differential gene expression analysis. As seen, at least 1.5M reads were ultimately quantified for each sample.

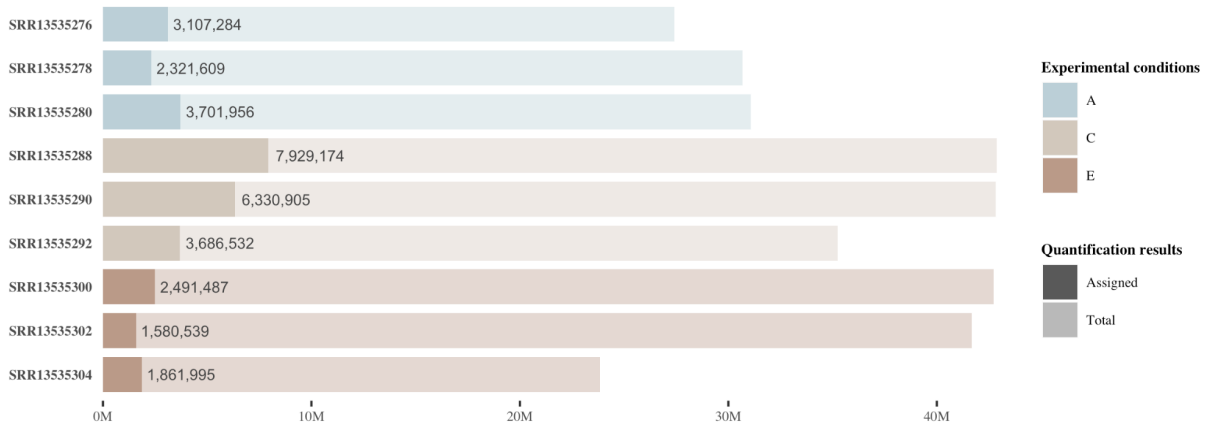

Figure 7. Quantification results, Paper #1

3.1.5 Differential gene expression results

For the A vs. C comparison (in space without gravity vs. in space with gravity), differential expression analysis identified 81 DE genes (9 upregulated and 72 downregulated). Using the paper's raw counts data to run the same analysis, 67 DE genes (7 upregulated and 60 downregulated) were identified. This is compared to the 61 total DE genes (9 upregulated and 52 downregulated) that the paper reported in the text. Because the paper only reported the number and not names of the DE genes they found, the 67 DE genes identified from using the paper's raw counts data were compared to the 81 identified by this project. As seen in Figure 8, the project identified 56 of the paper's 67 DE genes (83.6%), in addition to 25 additional DE genes.

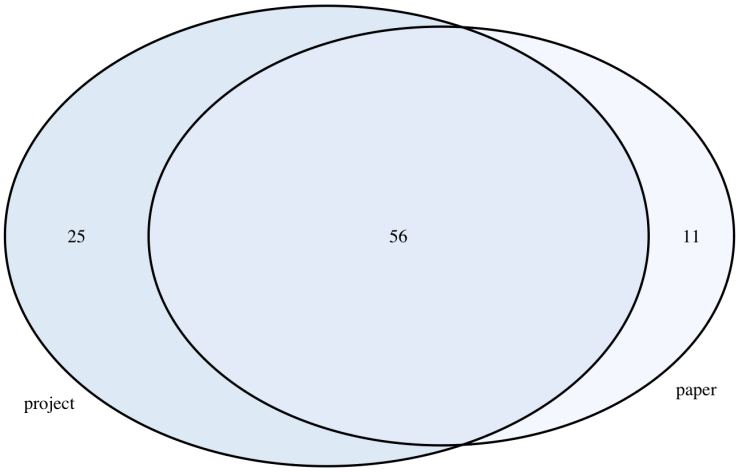

Figure 8. DE genes identified by project vs. paper for A vs. C comparison, Paper #1

For the A vs. E comparison (in space without gravity vs. on land), the project identified 496 DE genes (312 upregulated and 184 downregulated). Using the paper's raw counts data, 343 DE genes (239 upregulated and 104 downregulated) were identified. This is compared to the 343 total DE genes (230 upregulated and 113 downregulated) reported in the text. As seen in Figure 9, the project identified 295 of the 343 DE genes identified from using the paper's raw counts data (86.0%), in addition to having 201 additional DE genes.

Figure 9. DE genes identified by project vs. paper for A vs. E comparison, Paper #1

In general, the project was able to identify a majority of the same DE genes identified from using the paper's raw counts data. The project also seemed to identify a greater number of DE genes using the same significance level of 0.05 and LFC of greater than 1 criteria.

3.1.6 Gene set enrichment and pathway results

For the A vs. C comparison (in space without gravity vs. in space with gravity), GSEA results from using the paper's counts data and the project are largely similar. Figure 10 shows the top 10 most enriched and top 10 most depleted Hallmark gene sets from the project (top) and paper (bottom). Common enriched gene sets include the epithelial mesenchymal transition pathway at the top and depleted gene sets include protein secretion at the bottom.

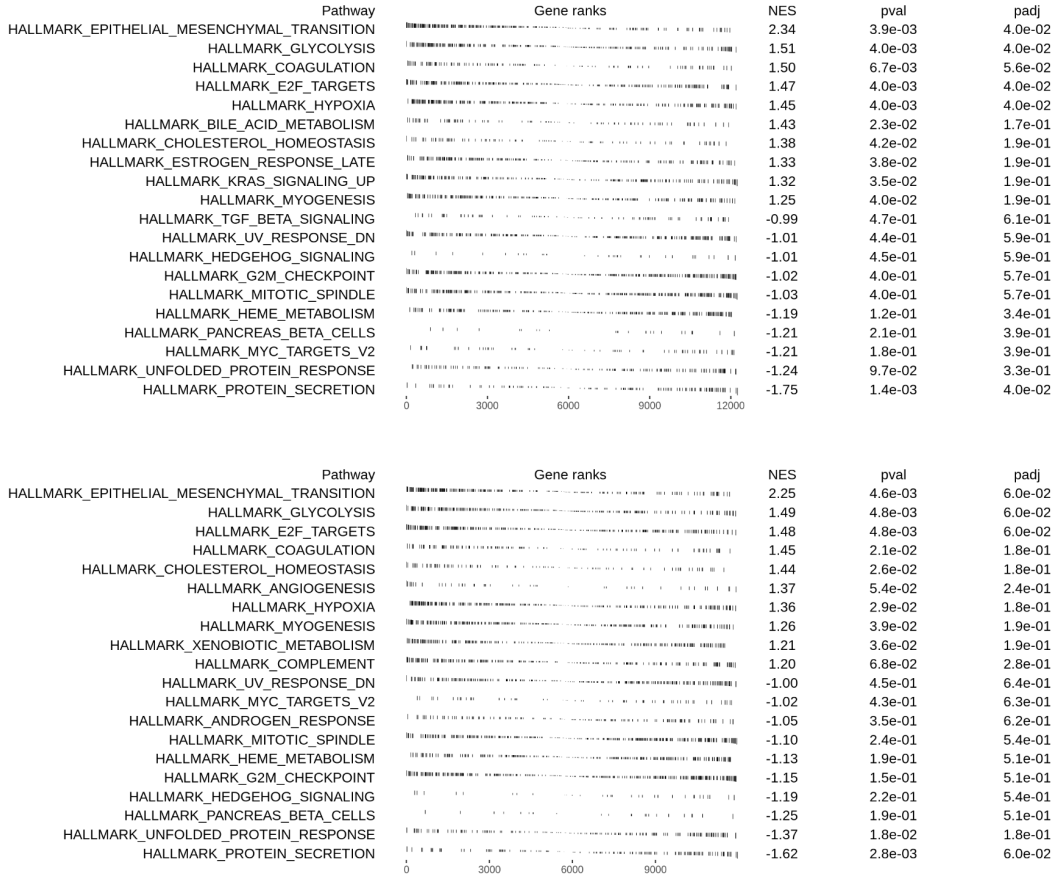

Figure 10. Top Hallmark gene sets identified by project (top) vs. paper (bottom) for A vs. C comparison, Paper #1

For the A vs. E comparison (in space without gravity vs. on land), both the project and paper data identified the myogenesis pathway as being significantly depleted, with an adjusted p-value of well below 0.25, as seen in Figure 11. Myogenesis is defined as the formation of skeletal muscles. This suggests that muscle atrophy could be an adverse effect of space, which is widely cited in literature. Interestingly, the results also show the epithelial mesenchymal transition pathway being depleted and the protein secretion being enriched, which is the opposite of what was observed for the A vs. C comparison.

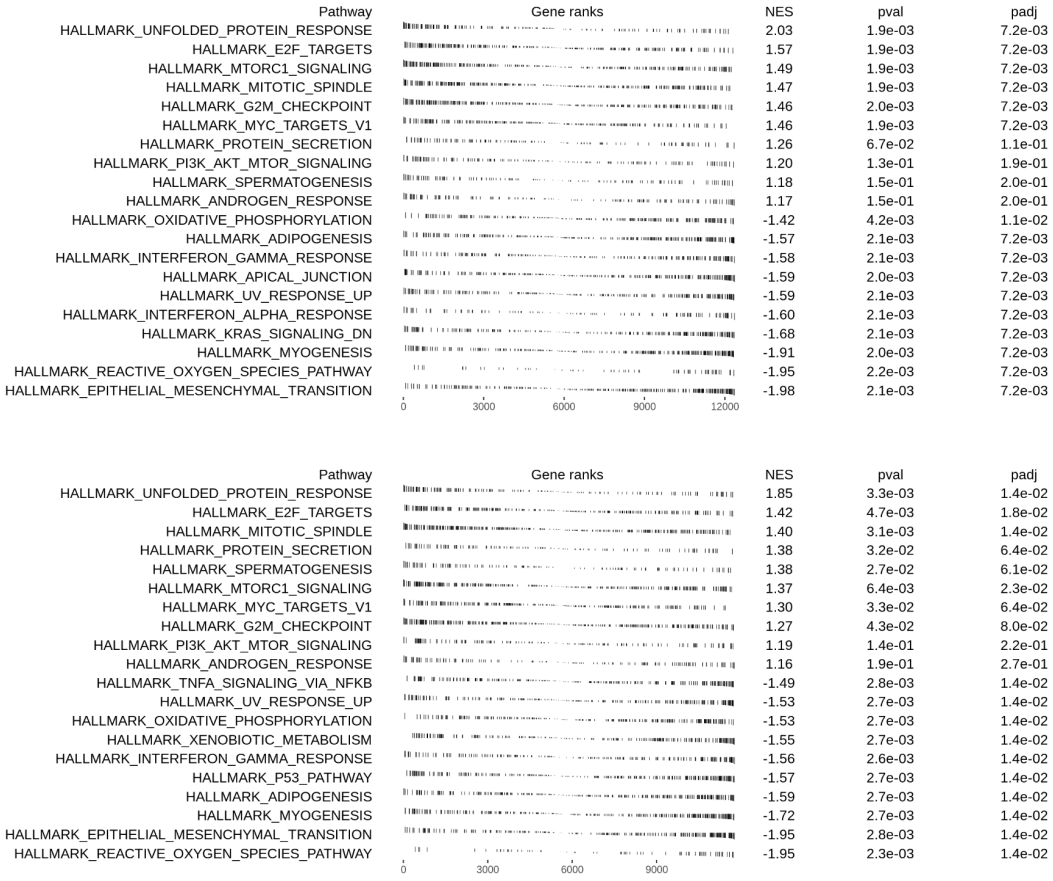

Figure 11. Top Hallmark gene sets identified by project (top) vs. paper (bottom) for A vs. E comparison, Paper #1

3.2 Paper #2

Raw RNA-seq data from six samples were downloaded for the second paper by Zhang et al. The total number of paired-end reads as well as the experimental condition for each sample is shown in Figure 12.

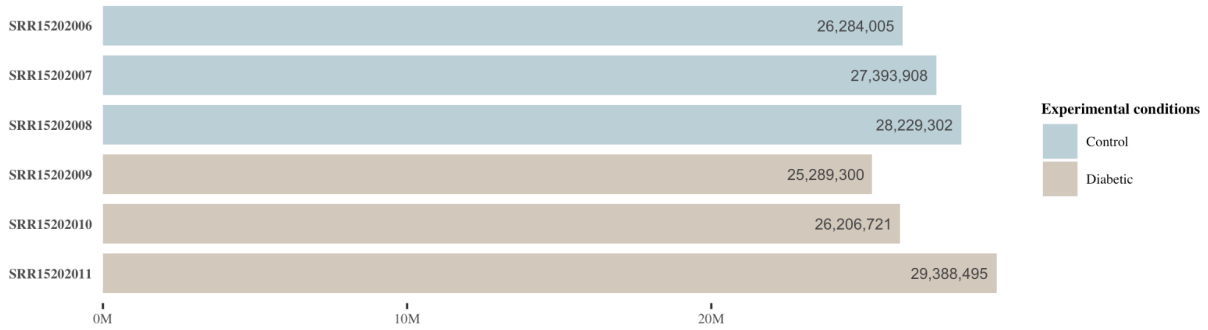

Figure 12. Raw read counts per sample, Paper #2

3.2.1 Trimming results

For the second paper's data, FastQC identified the adapter sequence as Illumina Universal Adapter, and this was successfully removed using Trimmomatic, as seen in Figure 13. After removing low-quality sequences and short sequences, over 95% of the reads still remained in all samples, as seen in Figure 14.

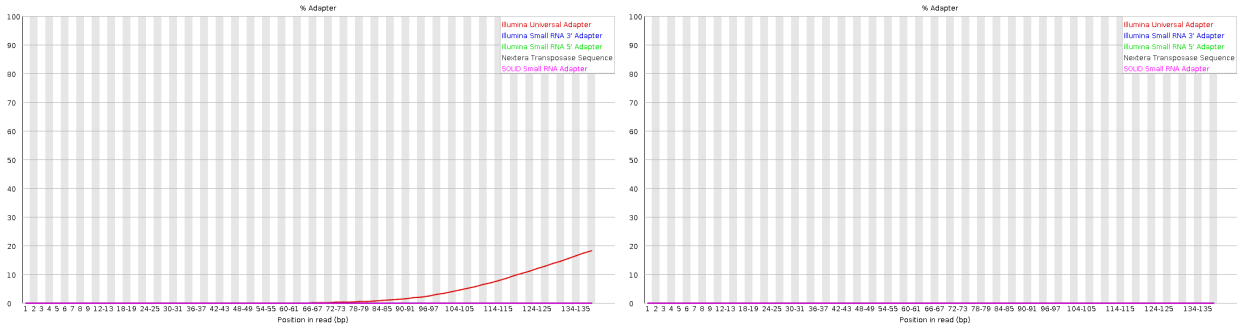

Figure 13. Adapter content before and after trimming, Paper #2

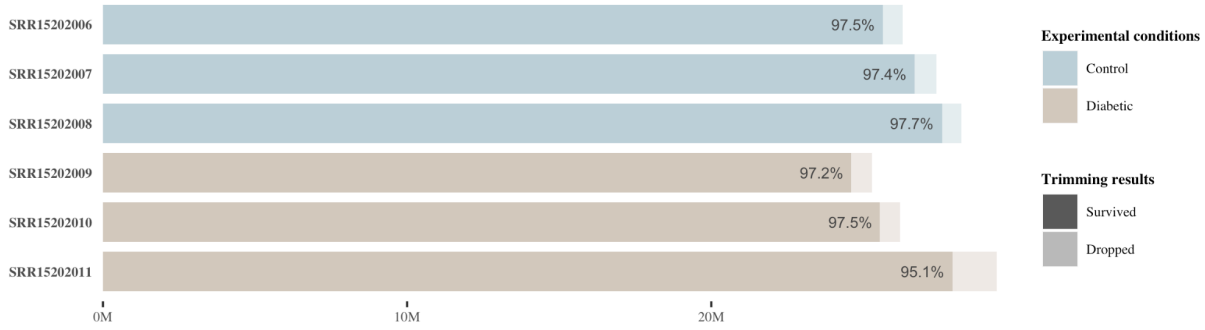

Figure 14. Trimming results, Paper #2

3.2.2 Mapping results

As seen in Figure 15, over 75% of reads that survived the trimming step in each sample were uniquely mapped to the reference genome. There are relatively low proportions of multi-mapped and unmapped reads, suggesting that the quality of the reads is high with few contaminations.

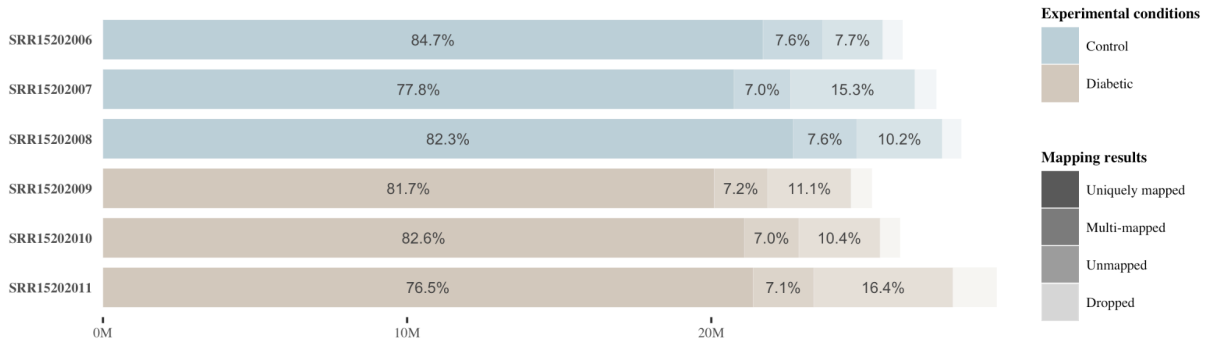

Figure 15. Mapping results, Paper #2

3.2.3 Deduplication results

Figure 16 shows the percentage of mapped reads that survived after removing duplicates. Generally, less than half of the reads were removed as duplicates.

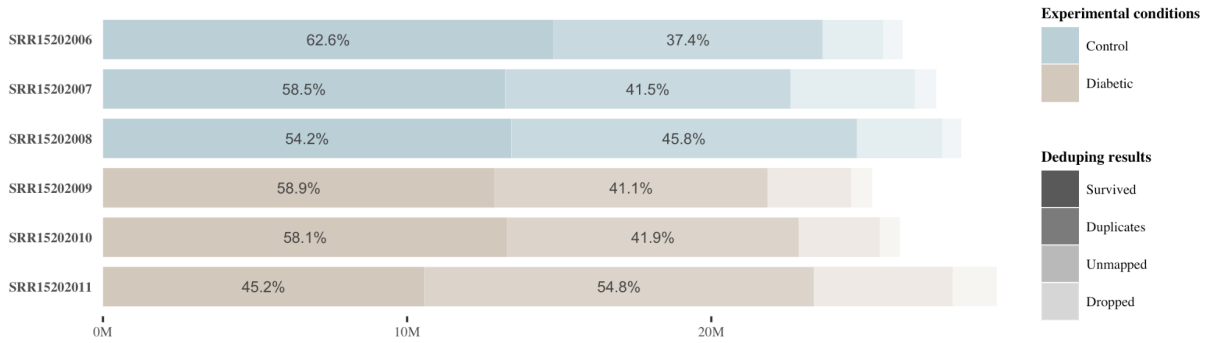

Figure 16. Deduplication results, Paper #2

3.2.4 Quantification results

Figure 17 shows the final number of read pairs that were assigned to a gene by featureCounts, as compared to the original number of raw reads that each sample started with. At least 8M reads were ultimately quantified for each sample.

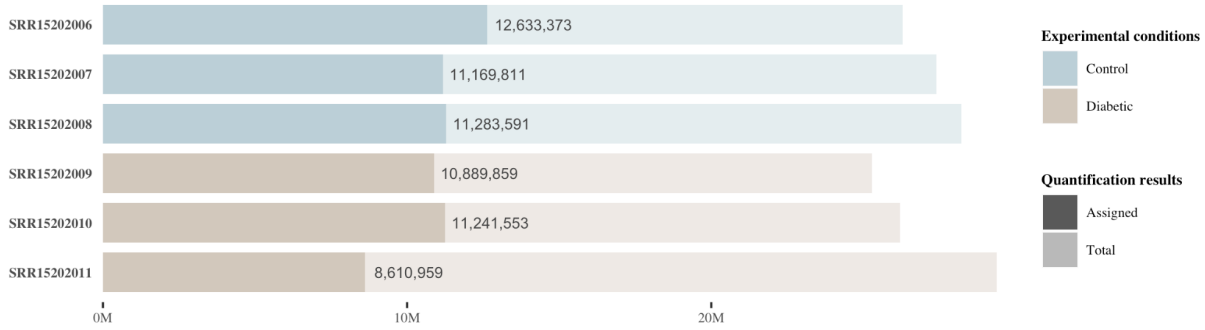

Figure 17. Quantification results, Paper #2

3.2.5 Differential gene expression results

267 DE genes (157 upregulated and 110 downregulated) were identified in the project, compared to the 186 DE genes (94 upregulated and 92 downregulated) reported in the paper. The paper provided a list of the DE genes they identified as part of their supplementary data. Figures 18 and 19 show the overlap between the upregulated and downregulated DE genes identified by the project and the paper.

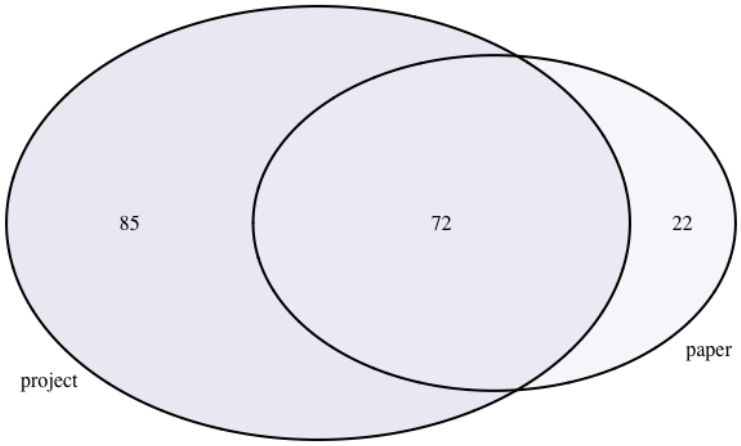

Figure 18. Upregulated genes identified by project vs. paper, Paper #2

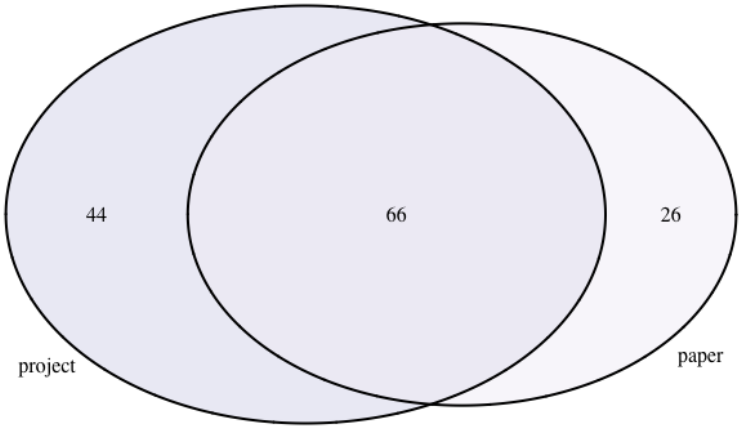

Figure 19. Downregulated genes identified by project vs. paper, Paper #2

As seen, the project was able to identify 72 of the paper's 94 upregulated genes (76.6%) and 66 of the paper's 92 downregulated genes (71.7%). In addition, the project identified additional genes as being differentially expressed. One thing to note is that the paper provided the list of genes by their gene symbol, whereas the project output used Ensembl gene ID. Conversions between the two were obtained using biomaRt, but they may be missing for some genes, which can contribute to some of the lack of overlap.

Figure 20 shows the volcano plot, which highlights the DE genes in color. As mentioned previously, an adjusted p-value of 0.05 and LFC threshold of 0.585 were used to identify DE genes, which can be seen in the plot. The labelled genes – *Sod3*, *Igf2*, *Ctgf*, and *Hmga2* – are ones that are identified in existing literature to be upregulated or downregulated in diabetic patients. The findings are consistent between the paper (left) and project (right) results.

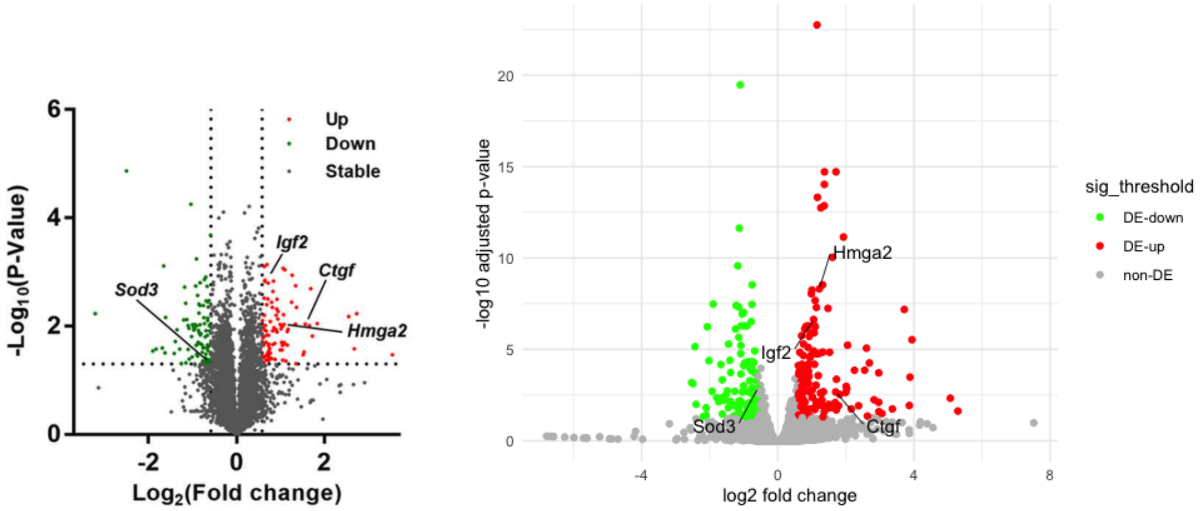

Figure 20. Volcano plot for paper (left) vs. project (right), Paper #2

3.2.6 Gene set enrichment and pathway results

GSEA was performed, and findings are also consistent with what was reported in the paper. For example, Hallmark pathways involved in immune response are enriched, including the inflammatory response, as seen in Figure 21. Glycolysis was also heavily depleted, which aligns with the fact that the treatment mice were diabetic. Among the GO Biological Process (BP) gene sets shown in Figure 22, the DE genes appear to be involved in pathways related to the immune response as well as epithelial regeneration and carbohydrate metabolism.

Pathway	Gene ranks	NES	pval	padj
HALLMARK_INTERFERON_ALPHA_RESPONSE		3.06	1.9e-03	7.0e-03
HALLMARK_INTERFERON_GAMMA_RESPONSE		2.75	1.9e-03	7.0e-03
HALLMARK_TNFA_SIGNALING_VIA_NFKB		1.93	1.9e-03	7.0e-03
HALLMARK_INFLAMMATORY_RESPONSE		1.81	1.9e-03	7.0e-03
HALLMARK_IL6_JAK_STAT3_SIGNALING		1.73	2.0e-03	7.0e-03
HALLMARK_ALLOGRAFT_REJECTION		1.58	1.9e-03	7.0e-03
HALLMARK_MYC_TARGETS_V2		1.51	1.4e-02	4.0e-02
HALLMARK_TGF_BETA_SIGNALING		1.40	3.9e-02	7.2e-02
HALLMARK_APOPTOSIS		1.36	2.7e-02	5.7e-02
HALLMARK_KRAS_SIGNALING_UP		1.29	3.8e-02	7.2e-02
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION		-1.43	2.1e-03	7.0e-03
HALLMARK_BILE_ACID_METABOLISM		-1.51	8.0e-03	2.5e-02
HALLMARK_HEME_METABOLISM		-1.52	2.1e-03	7.0e-03
HALLMARK_ADIPOGENESIS		-1.64	2.1e-03	7.0e-03
HALLMARK_MTORC1_SIGNALING		-1.71	2.1e-03	7.0e-03
HALLMARK_G2M_CHECKPOINT		-1.78	2.1e-03	7.0e-03
HALLMARK_HYPOXIA		-1.91	2.1e-03	7.0e-03
HALLMARK_E2F_TARGETS		-1.92	2.1e-03	7.0e-03
HALLMARK_OXIDATIVE_PHOSPHORYLATION		-2.12	2.0e-03	7.0e-03
HALLMARK_GLYCOLYSIS		-2.13	2.1e-03	7.0e-03

Figure 21. Top Hallmark gene sets, Paper #2

Pathway	Gene ranks	NES	pval	padj
GO_cellular_response_to_interferon-beta		2.49	1.9e-03	8.0e-02
GO_defense_response_to_virus		2.44	1.8e-03	8.0e-02
GO_response_to_interferon-beta		2.42	1.9e-03	8.0e-02
GO_response_to_virus		2.39	1.8e-03	8.0e-02
GO_negative_regulation_of_viral_life_cycle		2.10	1.9e-03	8.0e-02
GO_regulation_of_nuclease_activity		2.07	1.9e-03	8.0e-02
GO_negative_regulation_of_viral_genome_replication		2.07	1.9e-03	8.0e-02
GO_regulation_of_defense_response_to_virus_by_host		2.05	1.8e-03	8.0e-02
GO_antigen_processing_and_presentation_of_peptide_antigen_via_MHC_class_I		2.05	1.9e-03	8.0e-02
GO_regulation_of_defense_response_to_virus		2.00	1.8e-03	8.0e-02
GO_epithelial_cell_proliferation		-1.37	2.3e-03	8.0e-02
GO_carbohydrate_derivative_biosynthetic_process		-1.44	2.4e-03	8.0e-02
GO_cellular_carbohydrate_metabolic_process		-1.46	2.2e-03	8.0e-02
GO_gland_development		-1.47	2.3e-03	8.0e-02
GO_carbohydrate_metabolic_process		-1.51	2.3e-03	8.0e-02
GO_nucleotide_metabolic_process		-1.51	2.3e-03	8.0e-02
GO_nucleoside_phosphate_metabolic_process		-1.52	2.3e-03	8.0e-02
GO_ribose_phosphate_metabolic_process		-1.53	2.3e-03	8.0e-02
GO_ribonucleotide_metabolic_process		-1.55	2.3e-03	8.0e-02
GO_nucleobase-containing_small_molecule_metabolic_process		-1.56	2.3e-03	8.0e-02

Figure 22. Top GO Biological Process (BP) gene sets, Paper #2

IV. DISCUSSION

4.1 Removing duplicates

DE analysis was performed with and without removing duplicate reads. When duplicates are not removed, more DE genes tend to be identified, which naturally leads to more overlap with the papers' results. However, the increase in overlap is minimal compared to the number of additional DE genes the project identified that the paper did not commonly identify. As a more conservative approach, duplicates were removed as a way to reduce potential false positives.

4.2 Interacting factors

As seen in the GSEA results for the first paper, some results are contradicting between the pairwise comparisons. Specifically, certain pathways enriched in the A vs. C comparison are depleted in the A vs. E comparison, and vice versa. The first paper reported this phenomenon as well, explaining that the effects of microgravity and radiation may be opposing. This shows that it may be helpful to study multiple factors at the same time to see how they interact, such as the effects of microgravity in combination with different flow rates.

4.3 Pipeline validation

Data from two papers were used to test and validate the pipeline in this project. Broad findings and results were consistent between the project and paper data, however there are some discrepancies, which can be caused by differences in the tools and settings used along each step in the pipeline. In a real experiment, it may be useful to try out various tools to see how the end results compare and identify consistencies between the runs. The current mRNA-sequencing pipeline is designed to be flexible to swap out tools and versions of tools at each step, as well as adjust parameters using the configuration file.

4.4 Pipeline runtime

The pipeline also fully takes advantage of Snakemake and Slurm to produce a workflow that can be run in parallel for optimal efficiency. Figure 23 shows the time it takes to run the pipeline for the first (top) and second (bottom) papers. To make them more comparable, the maximum number of jobs to be run in parallel for each paper was limited to the number of samples they have, or 9 jobs for the first paper and 6 for the second. All steps were run with the default of 1 CPU core.

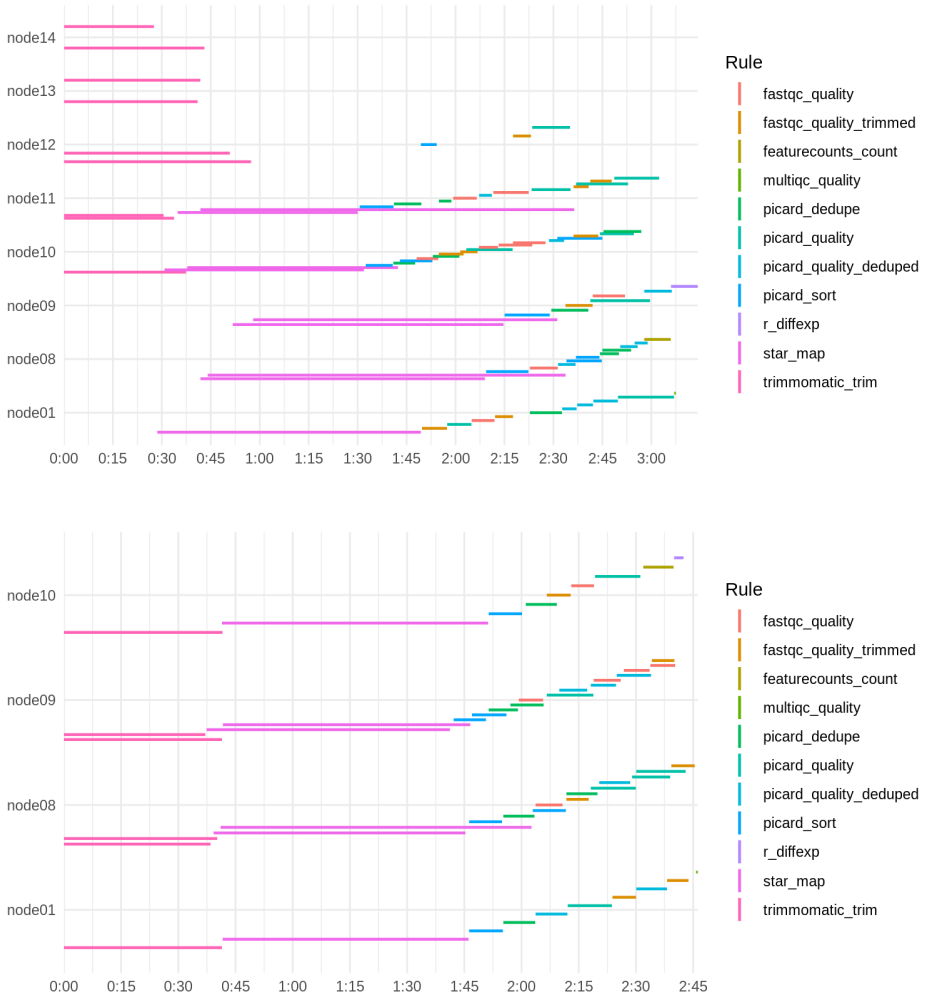

Figure 23. Pipeline runtime for Paper #1 (top) and Paper #2 (bottom), 1 CPU core per rule

As seen, the first paper took a little over three hours and the second paper took a little under three hours to run. It makes sense that the first paper would take longer given that its samples contain more reads. Being able to run the samples in parallel already significantly reduced the total runtime, as running each step for each sample sequentially could take upwards towards a full day to run.

From Figure 23, it can also be seen that the trimming and mapping steps take the longest time to run. Luckily, both Trimmomatic and STAR support multithreading, so more CPU cores can be allocated to them in the pipeline to speed up the runtime.

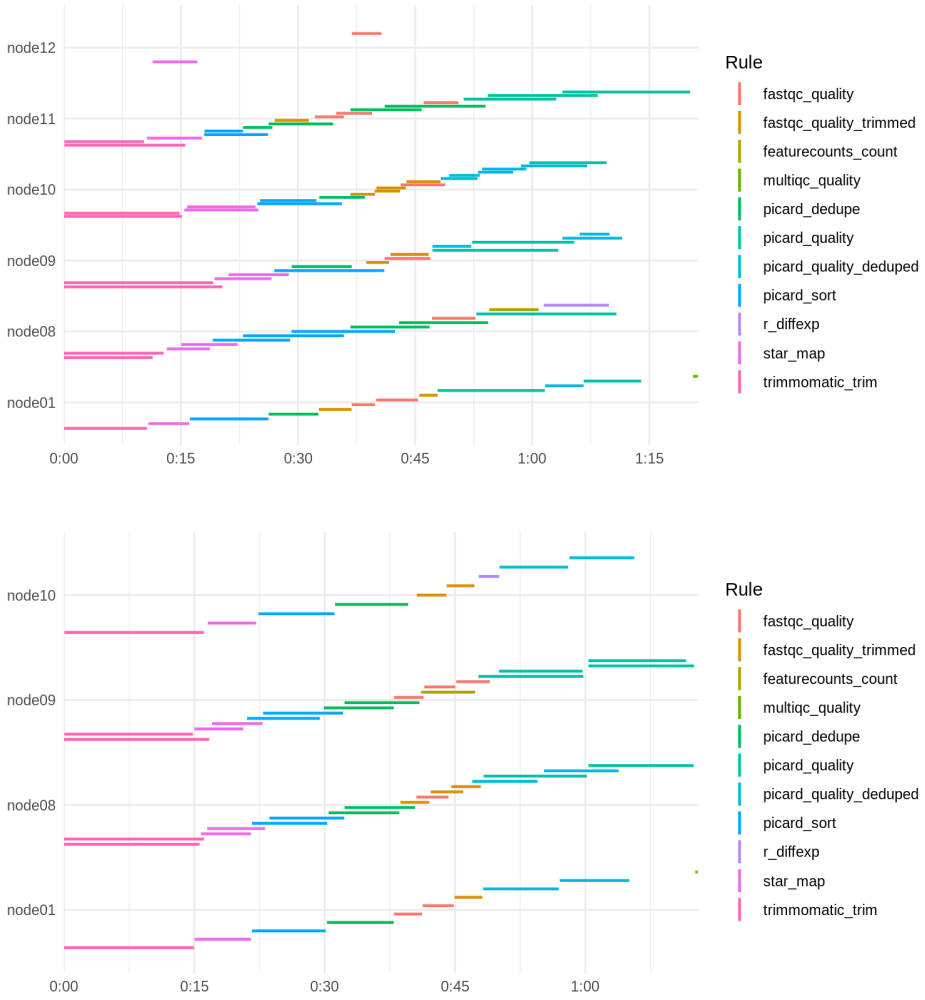

Figure 24. Pipeline runtime for Paper #1 (top) and Paper #2 (bottom), multithreading

Figure 24 shows the resulting pipeline runtime when more cores were used for the tools that support multithreading. Specifically, two cores were allocated to run FastQC, and 14 cores each were allocated to run Trimmomatic, STAR, and featureCounts. As seen, this helped cut down the total runtime in half.

4.5 Code and data availability

All code and documentation of the data used can be found in the GitHub repository here. Select outputs from the pipeline for the two papers' data, including the raw counts file and all generated reports, can be found here.

REFERENCES

[1] K. Marshall-Goebel et al., "Assessment of jugular venous blood flow stasis and thrombosis during spaceflight," *JAMA Network Open*, vol. 2, no. 11, Nov. 2019, doi: https://doi.org/10.1001/jamanetworkopen.2019.15011.

[2] J. W. Yau, H. Teoh, and S. Verma, "Endothelial cell control of thrombosis," *BMC cardiovascular disorders*, vol. 15, no. 130, Oct. 2015, doi: https://doi.org/10.1186/s12872-015-0124-z.

[3] N. Mackman, "New insights into the mechanisms of venous thrombosis," *The Journal of clinical investigation*, vol. 122, nol. 7, pp. 2331-2336, Jul. 2012, doi: https://doi.org/10.1172/JCI60229.

[4] S. Andrews, *FastQC: A Quality Control Tool for High Throughput Sequence Data*, 2010. [Online]. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[5] ecSeq Bioinformatics, *Why does the per base sequence quality decrease over the read in Illumina?*, Jan. 2017. [Online] Available: https://www.ecseq.com/support/ngs/why-does-the-sequence-quality-decrease-over-the-read-in-illumina.

[6] ecSeq Bioinformatics, *Trimming adapter sequences - is it necessary?*, Aug. 2016. [Online] Available: https://www.ecseq.com/support/ngs/trimming-adapter-sequences-is-it-necessary.

[7] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data," *Bioinformatics*, vol. 30, no. 15, pp. 2114-2120, Aug. 2014, doi: https://doi.org/10.1093/bioinformatics/btu170.

[8] A. Dobin, et al., "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics (Oxford, England)*, vol. 29, no. 1, pp. 15-21, Jan. 2013, doi: https://doi.org/10.1093/bioinformatics/bts635.

[9] S. G. Acinas, et al., "PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample," *Appl Environ Microbiol*, vol. 71, no. 12, pp. 8966-8969, Dec. 2005, doi: https://dx.doi.org/10.1128%2FAEM.71.12.8966-8969.2005.

[10] Broad Institute, *Picard Toolkit*, 2019. [Online]. Available: https://broadinstitute.github.io/picard/.

[11] Broad Institute, *MarkDuplicates (Picard)*, 2019. [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360037052812-MarkDuplicates-Picard-.

[12] Broad Institute, *SortSam (Picard)*, 2019. [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360036510732-SortSam-Picard-.

[13] Broad Institute, *CollectMultipleMetrics (Picard)*, 2019. [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360037594031-CollectMultipleMetrics-Picard-.

[14] Y. Liao, G. K. Smyth, and W. Shi, "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features," *Bioinformatics*, vol. 30, no. 7, pp. 923-930, Apr. 2014, doi: https://doi.org/10.1093/bioinformatics/btt656.

[15] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biology*, vol. 15, no. 550, Dec. 2014, doi: https://doi.org/10.1186/s13059-014-0550-8.

[16] hbctraining, *Differential gene expression workshop using Salmon counts*, 2020. [Online]. Available: https://hbctraining.github.io/DGE_workshop_salmon_online/.

[17] A. Subramanian, et al., "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15545-15550, Oct. 2005, doi: https://doi.org/10.1073/pnas.0506580102.

[18] GSEA, *Gene Set Enrichment Analysis (GSEA) User Guide*, Nov. 2019. [Online] Available: https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideFrame.html.

[19] G. Korotkevich, V. Sukhov, A. Sergushichev, "Fast gene set enrichment analysis," 2019, doi: https://doi.org/10.1101/060012.

[20] P. Ewels, et al., "MultiQC: summarize analysis results for multiple tools and samples in a single report," *Bioinformatics*, vol. 32, no. 19, pp. 3047-3048, Oct. 2016, doi: https://doi.org/10.1093/bioinformatics/btw354.

[21] J. Köster and S. Rahmann, "Snakemake–a scalable bioinformatics workflow engine," *Bioinformatics*, vol. 28, no. 19, pp. 2520-2522, Oct. 2012, doi: https://doi.org/10.1093/bioinformatics/bts480.

[22] A. B. Yoo, M. A. Jette, M. Grondona, "SLURM: Simple Linux Utility for Resource Management," in *Job Scheduling Strategies for Parallel Processing*. Berlin, Heidelberg: Springer, pp. 44-60, 2003, doi: https://doi.org/10.1007/10968987_3.

[23] G. G. Genchi, et al., "Cerium oxide nanoparticle administration to skeletal muscle cells under different gravity and radiation conditions," *ACS Appl. Mater. Interfaces*, vol. 13, no. 34, pp. 40200-40213, Sep. 2021, doi: https://doi.org/10.1021/acsami.1c14176.

[24] Y. Zhang, et al., "Transcriptional network analysis reveals the role of miR-223-5p during diabetic corneal epithelial regeneration," *Front Mol. Biosci*., vol. 8, Aug. 2021, doi: https://dx.doi.org/10.3389%2Ffmolb.2021.737472.

[25] timflutre, *Trimmomatic (adapters)*, 2015. [Online]. Available: https://github.com/timflutre/trimmomatic/tree/master/adapters.

[26] Broad Institute, *Phred-scaled quality scores*, 2021. [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360035531872-Phred-scaled-quality-scores.

[27] A. D. Yates, et al., "Ensembl 2020," *Nucleic Acids Research*, vol. 48, no. D1, pp. D682-D688, Jan. 2020, doi: https://doi.org/10.1093/nar/gkz966.

[28] W. J. Kent, et al., "The human genome browser at UCSC," *Genome Res.*, vol. 12, no. 6, pp. 996-1006, Jun. 2002, doi: https://doi.org/10.1101/gr.229102.

[29] L. Wang, S. Wang, W. Li, "RSeQC: quality control of RNA-seq experiments, " *Bioinformatics*, vol. 28, no. 16, pp. 2184-2185, Aug. 2012, doi: https://doi.org/10.1093/bioinformatics/bts356.

[30] Walter and Eliza Hall, *Mouse and Human Versions of the MSigDB in R Format*, 2021. [Online]. Available: https://bioinf.wehi.edu.au/MSigDB/index.html.

[31] S. Durinck, et al., "Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt," *Nat Protoc*, vol. 4, no. 8, pp. 1184-1191, Jul. 2009, doi: https://doi.org/10.1038/nprot.2009.97.